10 HYPOTHESIS TESTING

Objectives

After studying this chapter you should

- be able to define a null and alternative hypothesis;
- be able to calculate probabilities using an appropriate model to test a null hypothesis;
- be able to test for the mean based on a sample;
- understand when to use a one or two tailed test.

10.0 Introduction

One of the most important uses of statistics is to be able to make conclusions and test hypotheses. Your conclusions can never be absolutely sure, but you can quantify your measure of confidence in the result as you will see in this chapter.

Activity 1 Can you tell the difference?

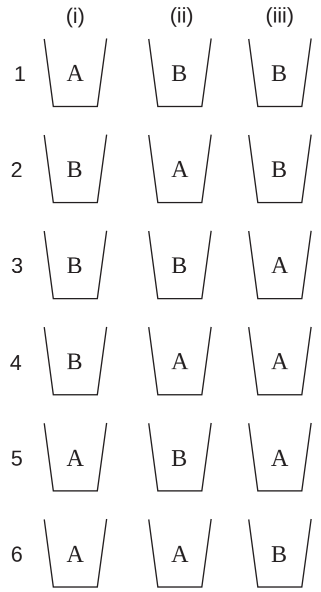
Can you tell HP Baked Beans from a supermarket brand? Can you tell Coca Cola from a supermarket brand?

You are going to set up an experiment to determine whether people really can tell the difference between two similar foods or drinks.

Each person taking part in the test is given 3 samples: two of one product and one of another (so that they may have two cups containing Coca Cola (say) and one cup containing a supermarket brand or vice versa).

Ask the subject to identify the sample which is different from the other two.

Note that there are six possible groups of samples and a die can be used to decide which grouping to give to each individual subject taking part.



	(i)	(ii)	(iii)
1	A	B	B
2	B	A	B
3	B	B	A
4	B	A	A
5	A	B	A
6	A	A	B

Plan the experiment carefully before you start. Write out a list showing the samples and order of presentation for all your subjects (about 12, say).

Ensure that your subjects take the test individually in quiet surroundings, free from odours. All 3 samples must be of the same size and temperature. If there are any differences in colour you can blindfold your subject. Record each person's answer. Count the number of subjects giving the correct answer. Subjects who are unable to detect any difference at all in the 3 samples must be left out of the analysis.

10.1 Forming a hypothesis

In any experiment you usually have your own hypothesis as to how the results will turn out.

However it is usual to set up a **null hypothesis** that states the opposite of what you want to prove. This can only then be abandoned in the face of overwhelming evidence, thus placing the onus of proof on you.

The null hypothesis H_0

For the activity above your null hypothesis is that subjects cannot tell the difference between the 3 samples and that they are guessing.

The alternative hypothesis H_1

This is your experimental hypothesis (or what you really wish to prove). For the activity above your alternative hypothesis is that subjects really can distinguish between the samples (or some of them can at least).

These hypotheses can be written in mathematical terms as :

$$H_0 : p = \frac{1}{3}$$

$$H_1 : p > \frac{1}{3}$$

Here p is the probability of success assuming that H_0 is true; that is, subjects cannot tell the difference and are randomly guessing.

In order to reject H_0 and adopt H_1, your **experimental** results will have to be ones which are very difficult to explain under the null hypothesis.

You can use the binomial distribution to calculate the probabilities of people achieving various results by guesswork. Here are the probabilities for all the possible results for 10 subjects.

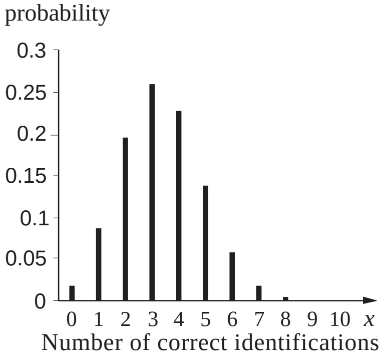
Number of people giving correct answer	Binomial probabilities	
0	$\binom{10}{0}\left(\frac{2}{3}\right)^{10}$	$= 0.0173$
1	$\binom{10}{1}\left(\frac{2}{3}\right)^{9}\left(\frac{1}{3}\right)$	$= 0.0867$
2	$\binom{10}{2}\left(\frac{2}{3}\right)^{8}\left(\frac{1}{3}\right)^{2}$	$= 0.1951$
3	$\binom{10}{3}\left(\frac{2}{3}\right)^{7}\left(\frac{1}{3}\right)^{3}$	$= 0.2601$
4	$\binom{10}{4}\left(\frac{2}{3}\right)^{6}\left(\frac{1}{3}\right)^{4}$	$= 0.2276$
5	$\binom{10}{5}\left(\frac{2}{3}\right)^{5}\left(\frac{1}{3}\right)^{5}$	$= 0.1366$
6	$\binom{10}{6}\left(\frac{2}{3}\right)^{4}\left(\frac{1}{3}\right)^{6}$	$= 0.0569$
7	$\binom{10}{7}\left(\frac{2}{3}\right)^{3}\left(\frac{1}{3}\right)^{7}$	$= 0.0163$
8	$\binom{10}{8}\left(\frac{2}{3}\right)^{2}\left(\frac{1}{3}\right)^{8}$	$= 0.0030$
9	$\binom{10}{9}\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{9}$	$= 0.0003$
10	$\binom{10}{10}\left(\frac{1}{3}\right)^{10}$	$= 0.00002$

As the probability of 10 people guessing correctly is so small, if this actually happened you would be much more inclined to believe that they can actually tell the difference between the samples. So in this case it would be more rational to reject H_0, because the explanation offered by H_1 is more plausible.

How many of the other possible results are not easily explained by H_o (and so better explained by H_1)?

Under H_0 the probability of :

10 correct guesses is	0.00002
9 or 10 correct guesses	0.00032
8, 9 or 10 correct guesses	0.00332
7, 8, 9 or 10 correct guesses	0.01962
6, 7, 8, 9 or 10 correct guesses	0.07652



Note that if you adopt 9 correct as a 'significant' result you must include the probability for 10 as well (because 10 is actually a 'better' result than 9). Similarly with 8 you must include the probabilities for 9 or 10 correct and so on.

In scientific experiments, it is usual to take results with probabilities of 0.05 (5%) or less as convincing evidence for rejecting the null hypothesis.

If 10 subjects take the taste test then you will conclude that they *can* tell the difference between the samples if 7 or more of them make correct identifications.

Activity 2

Use a similar analysis to test your hypothesis in Activity 1.

Exercise 10A

1. A woman who claims to be able to tell margarine from butter correctly picks the 'odd' sample out of the 3 presented, for 5 out of 7 trials. Is this sufficient evidence to back up her claim?
2. A company has 40% women employees, yet of the 10 section heads, only 2 are women. Is this evidence of discrimination against women?
3. A subject takes a test for ESP (extra-sensory perception) in which he has to identify the suit of a playing card held by the experimenter. (The experimenter can see the card, but the subject cannot.) For 10 cards he makes 7 correct identifications. Is this evidence of ESP?

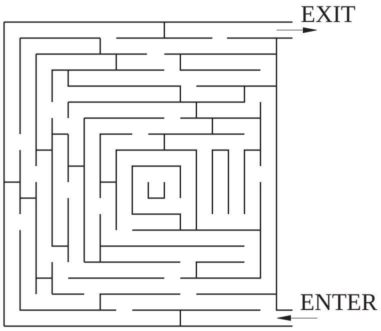
*10.2 The sign test

Another important use of hypothesis testing is to find out if, for a particular situation, you improve with practice.

Activity 3 Improve with practice?

(a) Maze

Time yourself finding your way through the maze shown opposite. Then have another try. Are you faster at the second attempt?



(b) Reaction times

Use a reaction ruler to find your reaction time. Record your first result and then your sixth (after a period of practice).

Analysing results

Suppose that 7 students took the maze test. Here are their times in seconds.

First try	Second try	Improvement?
9.0	3.5	√
6.7	4.0	√
5.8	2.6	√
8.3	4.6	√
5.1	5.4	X
4.9	3.7	√
9.2	5.7	√

Out of 7 subjects, 6 have improved, but could this result have occurred by chance? You can set up a hypothesis test in a similar way to the method used in the previous section.

Null hypothesis H_0

The null hypothesis is that any improvement or deterioration in times is quite random and that both are equally likely.

$$p(\text{improvement}) = \frac{1}{2}$$

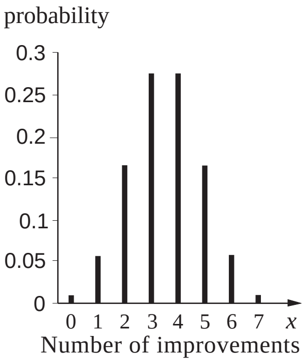
or $$p(\text{deterioration}) = \frac{1}{2}$$

Alternative hypothesis H_1

In this situation you expect people to improve with practice so the alternative hypothesis is that

$$p(\text{improvement}) > \frac{1}{2}$$

In order to analyse the experimental results ignore any students who manage to achieve identical times in both trials. Their results actually do not affect your belief in the null hypothesis either way. If there are n students with non-zero differences in times, the binomial distribution can be used as a model to generate probabilities under the null hypothesis. If X is the random variable 'number of positive differences', then $X \sim B(7, \frac{1}{2})$.



This gives the table of probabilities shown below

Number of positive differences	Probability
0	$\binom{7}{0}\left(\frac{1}{2}\right)^7 = 0.0078$
1	$\binom{7}{1}\left(\frac{1}{2}\right)^7 = 0.0547$
2	$\binom{7}{2}\left(\frac{1}{2}\right)^7 = 0.1641$
3	$\binom{7}{3}\left(\frac{1}{2}\right)^7 = 0.2734$
4	$\binom{7}{4}\left(\frac{1}{2}\right)^7 = 0.2734$
5	$\binom{7}{5}\left(\frac{1}{2}\right)^7 = 0.1641$
6	$\binom{7}{6}\left(\frac{1}{2}\right)^7 = 0.0547$
7	$\binom{7}{7}\left(\frac{1}{2}\right)^7 = 0.0078$

The binomial probabilities have been calculated according to $H_0 : X \sim B(7, \tfrac{1}{2})$. Under H_0, the probability of :

7 improvements is	0.0078
6 or 7 improvements is	0.0625

If you adhere to a 5% level of significance (0.05 probability of rejecting H_0 when it may be true), the result of 6 improvements is actually not sufficient grounds for rejecting H_0. This method of hypothesis testing is called the **sign test**.

Activity 4

Follow through the method outlined in the previous section to analyse the results of your experiments in Activity 3.

*Exercise 10B

In all questions, assume a 5% level of significance.

1. A group of students undertook an intensive six week training programme, with a view to improving their times for swimming 25 metres breast-stroke. Here are their times measured before and after the training programme. Have they improved significantly?

25m breast stroke times in seconds

student	A	B	C	D	E	F	G	H
before programme	26.7	22.7	18.4	27.3	19.8	20.2	25.2	29.8
after programme	22.5	20.1	18.9	24.8	19.5	20.9	24.0	24.0

2. Twelve young children (6-year-olds) were given a simple jigsaw puzzle to complete. The times they took were measured on their first and second attempts. Did they improve significantly?

jigsaw puzzle times in seconds

child	1	2	3	4	5	6	7	8	9	10	11	12
first attempt	143	43	271	63	232	51	109	156	304	198	83	115
second attempt	58	45	190	49	178	58	73	127	351	170	74	97

3. A group of 9 children wanted to see whether the amount of air in their bicycle tyres made a difference in how easy it was to pedal their bikes. They decided to ride a particular route under two different conditions : once with a tyre pressure of 40 pounds per square inch (psi) and once with 65 psi. (The order in which they did this was to be decided by tossing a coin.) The time it took (in minutes) for each circuit was :

40 psi	34	54	23	67	46	35	49	51	27
65 psi	32	45	21	63	37	40	51	39	23

Are the children significantly faster with the higher pressure tyres?

4. A group of engineering students run a test to see whether cars will get as many mpg on lead-free petrol as on 4-star petrol.

	car									
	A	B	C	D	E	F	G	H	I	J
lead-free	15	23	21	35	42	28	19	32	31	24
4-star	18	21	25	34	47	30	19	27	34	20

Does 4-star petrol give significantly better results?

5. A personnel director of a large company would like to know whether it will take less time to type a standard monthly report on a word processor or on a standard electric typewriter. A random sample of 7 secretaries was selected and the amount of typing time recorded in hours.

	secretary						
	A	B	C	D	E	F	G
electric typewriter	7.0	7.4	7.8	6.7	6.1	8.1	7.5
word processor	6.3	7.5	6.8	6.0	5.3	7.4	7.2

Are the secretaries significantly faster using the word processors?

10.3 Hypothesis testing for a mean

You can now extend the ideas, introduced in earlier sections, to the testing of a hypothesis about the mean of a sample. There are two cases to consider, firstly tests for the mean based on a sample from a normal distribution with known variance, and secondly tests based on a large sample from an unspecified distribution.

Example

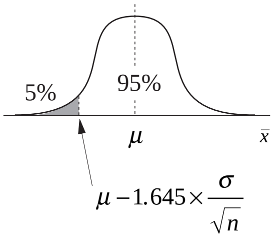
Afzal weighs the contents of 50 more packets of crisps and finds that the mean weight of his sample is 24.7 g. The weight stated on the packet is 25 g and the manufacturers claim that the weights are normally distributed with standard deviation 1 g. Can Afzal justifiably complain that these packets are underweight?

Solution

For this problem

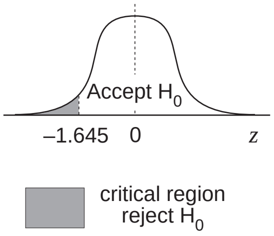
$$H_0 : M = 25\text{g}$$

$$H_1 : M < 25\text{g}$$



As Afzal suspects that the crisps are underweight he will reject the null hypothesis for unusually low values of $\overline{X}$.   The critical region consists of these values at the extreme left hand end of the distribution of $\overline{X}$, which have a 5% probability in total. (This is called a **one tailed test**.)

The critical value of z, which can be found from normal distribution tables, is -1.645.



Under H_0, $\overline{X} \sim N\left(25, \frac{1}{50}\right)$

Now the test statistic is

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\text{standard error}} \\ &= \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \\ &= \frac{24.7 - 25}{\left(\frac{1}{\sqrt{50}}\right)} \\ &= -2.12 \end{aligned}$$

This value of z is significant as it is less than the critical value, –1.645, and falls in the critical region for unusual values of $\overline{X}$. As it is extremely unlikely under H_0 (and is better explained by H_1) you can reject H_0. Afzal's results are such that he has good cause to complain to the manufacturers!

Example 2

A school dentist regularly inspects the teeth of children in their last year at primary school. She keeps records of the number of decayed teeth for these 11-year-old children in her area. Over a number of years, she has found that the number of decayed teeth was approximately normally distributed with mean 3.4 and standard deviation 2.1.

She visits just one middle school in her rounds. The class of 28 12-year-olds at that school have a mean of 3.0 decayed teeth. Is there any significant difference between this group and her usual 11-year-old patients?

Solution

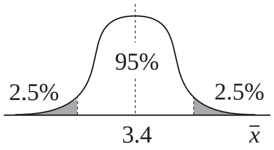
For this problem

$$\mathrm{H}_0 : M = 3.4$$

$$\mathrm{H}_1 : M \neq 3.4$$

The dentist has no reason to suspect either a higher or lower figure for the mean for 12-year-olds. (Children at this age may still be losing milk teeth) so the alternative hypothesis is non directional and a **two tailed test** is used.

The critical region (consisting of unusual results with low probabilities) is split, with $2\frac{1}{2}\%$ at both extremes of the distribution. The critical values of z are ± 1.96, from normal distribution tables.

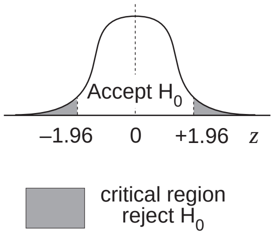


As before, the test statistic is

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$= \frac{3.0 - 3.4}{\left(\frac{2.1}{\sqrt{28}}\right)}$$

$$= -1$$



This value of z is not significant. It lies well within the main body of the distribution for $\bar{X}$. You must accept H_0 and conclude that the result for the 12-year-olds is not unusual.

When the distribution is unknown and the variance, σ^2, unknown, you have to use the Central Limit Theorem which states that the distribution of the sample means is normally distributed,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Since σ^2 is unknown, the estimate

$$\hat{\sigma}^2 = \frac{ns^2}{n-1}$$

is used, when s^2 is the sample variance.

But for larger n,

$$\frac{n}{n-1} \approx 1$$

and so $\hat{\sigma}^2 = s^2$, and you use the test statistics

$$z = \frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}.$$

Example

A manufacturer claims that the average life of their electric light bulbs is 2000 hours. A random sample of 64 bulbs is tested and the life, x, in hours recorded. The results obtained are as follows:

$$\sum x = 127\ 808 \qquad \sum(\bar{x}-x)^2 = 9694.6$$

Is there sufficient evidence, at the 1% level, that the manufacturer is over estimating the length of the life of the light bulbs?

Solution

From the sample

$$\bar{x} = \frac{\sum x}{n} = \frac{127\ 808}{64} = 1997$$

$$s^2 = \frac{\sum(\bar{x}-x)^2}{n} = \frac{9694.6}{64} = 151.48$$

giving the sample standard deviation as

$$s = 12.31.$$

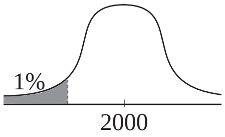
Let X be the random variable, the life (in hours) of a light bulb, so define

$$\mathrm{H}_0 : \mu = 2000$$

$\mathrm{H}_1 : \mu < 2000$ (assuming the manufacturer is over estimating the lifetime)

Assuming H_0,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \approx N\left(2000, \frac{151.48}{64}\right).$$



For a one tailed test at 1% significance level, the critical value of z is -2.33 (from normal distribution tables), and here

$$z = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

$$= \frac{1997 - 2000}{\left(\frac{12.31}{8}\right)}$$

$$= -1.95.$$

-1.95 is **not** inside the critical region so you conclude that, at 1% level, there is not sufficient evidence to reject H_0.

Exercise 10C

1. Explain, briefly, the roles of a null hypothesis, an alternative hypothesis and a level of significance in a statistical test, referring to your projects where possible.

   A shopkeeper complains that the average weight of chocolate bars of a certain type that he is buying from a wholesaler is less than the stated value of 8.50 g. The shopkeeper weighed 100 bars from a large delivery and found that their weights had a mean of 8.36 g and a standard deviation of 0.72 g. Using a 5% significance level, determine whether or not the shopkeeper is justified in his complaint. State clearly the null and alternative hypotheses that you are using, and express your conclusion in words.

   Obtain, to 2 decimal places, the limits of a 98% confidence interval for the mean weight of the chocolate bars in the shopkeeper's delivery.

2. At an early stage in analysing the marks scored by the large number of candidates in an examination paper, the Examination Board takes a random sample of 250 candidates and finds that the marks, x, of these candidates give $\sum x = 11\,872$ and $\sum x^2 = 646\ 193$. Calculate a 90% confidence interval for the population mean, μ, for this paper.

   Using the figures obtained in this sample, the null hypothesis $\mu = 49.5$ is tested against the alternative hypothesis $\mu < 49.5$ at the α% significance level. Determine the set of values of α for which the null hypothesis is rejected in favour of the alternative hypothesis.

   It is subsequently found that the population mean and standard deviation for the paper are 45.292 and 18.761 respectively. Find the probability of a random sample of size 250 giving a sample mean at least as high as the one found in the sample above.

10.4 Hypothesis testing summary

To summarise, note that:

The null hypothesis H_0 is an assertion that a parameter in a statistical model takes a **particular value**.

The alternative hypothesis H_1 expresses the way in which the value of a parameter may deviate from that specified in the null hypothesis.

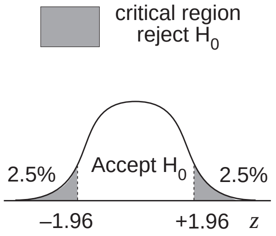
Critical region. A value of the test statistic is chosen so that it is very unlikely under H_0 and would be better explained by H_1. If the sample then generates a test statistic in this region defined by the critical value, H_0 will be rejected.

(a) **Two tailed tests**: You do not expect change in any particular direction.

$H_0 : M = k$, a particular value

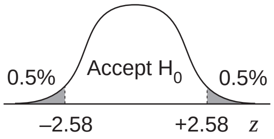
$H_1 : M \neq k$

Test statistic $z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$



Testing at the 5% level

The probability of incorrectly rejecting H_0 is 5%.



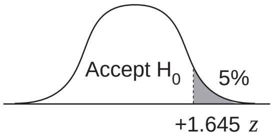
Testing at the 1% level

The probability of incorrectly rejecting H_0 is 1%.

(b) **One tailed tests**: You expect an increase

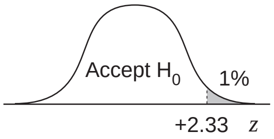
$H_0 : \mu = K$

$H_1 : \mu > K$



Testing at the 5% level

The probability of incorrectly rejecting H_0 is 5%.



Testing at the 1% level

The probability of incorrectly rejecting H_0 is 1%.

Similarly, if there are grounds for suspecting a decrease,

$H_0 : \mu = K$

$H_1 : \mu < K$

Note

(i) H_0 is the same for every test. It is H_1 which determines the position of the critical region.

(ii) It is always safer to use a two tailed test (unless you have very strong reasons to do otherwise).

Hypothesis testing method

1. Decide which is the variable under investigation.
2. Is it a discrete or a continuous variable?
3. What probability model can you use? (e.g. binomial, normal, uniform)
4. What is the null hypothesis? (H_0)
5. What is the alternative hypothesis? (H_1)
6. Sketch the distribution according to the null hypothesis.
7. Does the alternative hypothesis lead you to look for unusual values of x at one end of the distribution or both? (one or two tailed test?)
8. Is your result significant? (Does it lie in the critical region?)

10.5 Miscellaneous Exercises

1. A nutritionalist working for a babyfood manufacturer wishes to test whether a new variety of orange has a vitamin C content similar to the variety normally used by his company. The original variety of oranges has a mean vitamin C content of 110 milligrams and a standard deviation of 13 mg. His test results are (in mg)

   88, 109, 76, 136, 93, 101, 89, 115, 97, 92,

   106, 114, 109, 91, 94, 85, 87, 117, 105

   What are your conclusions? What assumptions did you need to make?

2. An engineer believes that her newly designed engine will save fuel. A large number of tests on engines of the old variety yielded a mean fuel consumption of 19.5 miles per gallon with standard deviation of 5.2. Fifteen new engines are tested, and give a mean fuel consumption of 21.6 miles per gallon. Is this a significant improvement?

3. A physiotherapist believes that exercise can slow down the ageing process. For the past few years she has been running an exercise class for a group of fourteen individuals whose average age is 50 years. Generally as a person ages, maximum oxygen consumption decreases. The national norm for maximum oxygen consumption in 50-year-old individuals is 30 millilitres per kilogram per minute with a standard deviation of 8.6. The mean for the members of the exercise class is 36 millilitres per kilogram per minute. Does the result support the physiotherapist's claim?

4. A coal merchant sells his coal in bags marked 50 kg. He claims that the mean mass is 50 kg with a standard deviation 1 kg. A suspicious weights and measures inspector has 60 of the bags weighed, and finds that their mean mass is 49.6 kg. Are the inspector's suspicions justified?

5. A sample of size 36 is taken from a population having mean μ and standard deviation 9; the sample mean is 47.4.

   Test the hypothesis $H_0 : \mu = 50$ against the alternative $H_1 : \mu < 50$ using the 5% level of significance.